Attorney Docket No.: 019404-001200US

PATENT APPLICATION

GENEALOGY INVESTIGATION AND DOCUMENTATION SYSTEMS AND METHODS

Inventor(s): Bennett Cookson, Jr., a citizen of the United States, residing at
439 North 950 East
Orem, UT 84097

Ken Boyer, a citizen of the United States, residing at
4430 West Windsor Street
Cedar Hills, UT 84062

James Mark Hamilton, a citizen of the United States, residing at
522 West 4630 North
Provo, UT 84604

Kendall J. Jefferson, a citizen of the United States, residing at
774 South 850 East
Orem, UT 84097

Daren Thayne, a citizen of the United States, residing at
95 East 1960 North
Orem, UT 84057

Michael J. Wolfgramm, a citizen of the United States, residing at
1480 East 860 South
Pleasand Grove, UT 84062

Assignee: MyFamily.com, Inc.
360 W 4800 N
Provo, UT 84604

Entity: Small business concern

TOWNSEND and TOWNSEND and CREW LLP
Two Embarcadero Center, Eighth Floor
San Francisco, California 94111-3834
Tel: 303-571-4000

GENEALOGY INVESTIGATION AND DOCUMENTATION SYSTEMS AND METHODS

CROSS-REFERENCES TO RELATED APPLICATIONS

[0001] This application is related to co-pending, commonly assigned and concurrently filed U.S. Patent Application No. _____, entitled, " PROVIDING ALTERNATIVES WITHIN A FAMILY TREE SYSTEMS AND METHODS " (Attorney Docket No. 019404-001400), by Bennett Cookson, Jr., *et al.*, and to co-pending, commonly assigned and concurrently filed U.S. Patent Application No. _____, entitled, " CORRELATING GENEALOGY RECORDS SYSTEMS AND METHODS " (Attorney Docket No. 019404-001500), by Bennett Cookson, Jr., *et al.*, the entire disclosure of each of which is herein incorporated by reference for all purposes.

BACKGROUND OF THE INVENTION

[0002] The present invention relates generally to genealogy and more particularly to computer-based genealogy investigation tools.

[0003] Genealogy is an enjoyable hobby to some and an important life's work to many. Whether for cultural, religious, recreational or other reasons, many people wish to trace their ancestry.

[0004] The process of genealogy investigation has evolved considerably over the years. In the past, the practice involved keeping notes in family bibles handed down through the generations, and many continue to do this today. Not very long ago, the process often required traveling to the hometowns of ancestors to pore over public records, newspapers, and the like at courthouses, libraries, and such. Once found, family information was written into journals and notebooks or onto index cards. Because of the geometric expansion of information with each generation, analyzing the information became a daunting task. The advent of computers, however, has created significant opportunities for improving and simplifying the process.

[0005] Many public records are now accessible using a computer and the Internet, thus allowing investigators to search electronically using keywords and such without having to travel to where the original records are kept. Additionally, several public and private efforts

to collect and catalog genealogy data have resulted in publicly accessible databases with much of the work already complete. Further still, some companies have produced commercial web sites where individuals can cooperate to extend a common family tree. Some examples of each include: <www.archives.gov>, the US National Archives and Record Administration website; <www.familysearch.org>, the LDS Church Family Search website; <www.ancestry.com>, the Ancestry.com website, which includes the Ancestry World Tree; <www.genealogy.com>, the Genealogy.com website, which (includes the World Family Tree); <www.ellisisland.org>, which includes immigration records; <www.interment.net>, which includes Cemeteries and Cemetery Records; <www.rootsweb.com>, which includes World Connect; <www.onegreatfamily.com>, the One Great Family website; <www.MyTrees.com>; and <www.GenCircles.com>. In fact, the process has become so popular that a standard data format has evolved.

[0006] GEDCOM (Genealogical Data Communication) is an industry standard data format for genealogical information. It uses a standard ASCII file format in which each line contains one data element. [A complete description of the GEDCOM file format is available at <www.gendex.com/gedcom55/55gcint.htm>, the content of which is entirely incorporated herein by reference for all purposes.] Many genealogy investigation services now collect and distribute data using the GEDCOM standard.

[0007] Despite the technological advances – or in some cases because of the technological advances – relating to genealogy, the activity remains ripe for improvement. One significant limitation that exists in many "open" genealogy investigation tools (*i.e.*, those that allow independent users to submit data), is a bias in favor of the information submitted by the most recent submitter. Because of the way data is related within these systems, data conflicts are difficult to resolve. The problem is rectified by allowing the latest submitter to overwrite conflicting data submitted by a previous user. This is but one example of the many limitations of presently-available genealogy investigation tools. Embodiments of the present invention address these and many other limitations.

BRIEF SUMMARY OF THE INVENTION

[0008] Embodiments of the present invention thus provide a method of creating a family tree. The method includes receiving genealogy data at a host computing system from at least one primary source and creating one or more node records and one or more link records using the genealogy data. The individual node records include at least name data and each

individual link record includes relationship data that represents a relationship between individual node records. The method also includes comparing individual node records and identifying pairs of records having similar data. For each identified pair of individual node records, the method includes comparing related individual node records and deciding based on predetermined criteria whether the identified pair of individual node records represent the same person. The method also includes consolidating the information from a plurality of records determined to represent the same person into a single person record. The method also includes receiving a request at the host computing system from a user computer to display a family tree and using the individual link records, the individual node records, and the single person records to create a data representation comprising the requested family tree. The method also includes sending the data representation to the user computer.

[0009] In some embodiments, the method includes using the genealogy data to create surname records. A surname record may include a surname and a number representing the number of times the corresponding surname is encountered in the genealogy data. The method may include using the surname records to partition the individual node records into groups prior to comparing the individual node records. Comparing individual node records and identifying pairs of records having similar names may include calculating a score representing the likelihood that the identified pair of individual node records represent the same person. Comparing related individual node records and deciding based on predetermined criteria whether the identified pair of individual node records represent the same person may include revising the score based on the comparison. The individual node records may span only a single generation or may span multiple generations. Receiving genealogy data from at least one source may include receiving genealogy data from a source such as the Ancestry World Tree system, a Social Security Death Index database, the World Family Tree system, a birth certificate database, a death certificate database, a marriage certificate database, an adoption database, a draft registration database, a veterans database, a military database, a property records database, a census database, a voter registration database, a phone database, an address database, a newspaper database, an immigration database, a family history records database, a local history records database, a business registration database, a motor vehicle database, and the like. Receiving genealogy data from at least one source may include receiving genealogy data as a GEDCOM file. Using the individual link records, the individual node records, and the single person records to create a file comprising the requested family tree may include including alternatives for relationships

for display to a user, in which case the method may include receiving a selection representing a user choice among the alternatives, using the selection to update the family tree, and storing the selection. In some embodiments the method includes receiving new information that changes the family tree and providing the user an opportunity to revise the selection. The method may include receiving information from a user. The information may include a digital picture, a text file, genealogy data, a user-entered text file, a sound file, a video file, any computer readable file, and the like, and storing the information. The information may be available to other users. The method may include receiving additional genealogy data that changes the family subsequent to sending the file to the user computer and notifying the user of the changes. Notifying the user may include sending the user an email, sending a file to the user upon the user accessing the host computing system, wherein the file comprises alternatives, displaying a notification to the user upon the user accessing the host computing system, and the like. The method may include receiving a request from the user computer to send more detailed information relating to the family tree subsequent to sending the file to the user computer, using the individual link records, the individual node records, and the single person records, to compile the more detailed information, and sending the more detailed information to the user computer.

[0010] In other embodiments the present invention provides a system for creating a family tree. The system includes a host computing system that includes means for receiving genealogy data from at least one primary source and means for sending information to a user computer. The host computer system is programmed to create one or more node records and one or more link records from received genealogy data. The individual node records include at least name data and each individual link record includes relationship data that represents a relationship between individual node records. The host computer system is also programmed to compare individual node records and identify pairs of records having similar data and for each identified pair of individual node records, compare related individual node records and decide based on predetermined criteria whether the identified pair of individual node records represent the same person. The host computer system is further programmed to consolidate the information from a plurality of records determined to represent the same person into a single person record and respond to a request from a user computer to display a family tree by using the individual link records, the individual node records, and the single person records to create a data representation comprising the requested family tree. The host computer is also programmed to send the data representation to the user computer.

[0011] In still other embodiments the present invention provides a method of creating a family tree that includes receiving data at a host computer system that defines a plurality of personas. The data includes one or more assertions for each persona and each persona represents a person. The method also includes storing each persona as a persona record and receiving a request at the host computer system from a user to provide a family tree. The request includes at least one assertion. The method also includes identifying an initial persona record and from the initial persona record, performing a relationship analysis to infer any relationships with other persona records using the assertions of the initial persona record and the other persona records. If a relationship is inferred, at least one relationship type is assigned to the relationship between the records. The method also includes using the persona records and the relationship types to construct a family tree and sending a file comprising at least a portion of the family tree to the user.

[0012] In still other embodiments the present invention provides a system for creating a family tree. The system includes a host computer system that is configured to receive data that defines a plurality of personas. The data includes one or more assertions for each persona and each persona represents a person. The host computer system is further configured to store each persona as a persona record and perform a relationship analysis to infer relationships among persona records using the assertions of the persona records. If a relationship is inferred, at least one relationship type is assigned to the relationship between the records. The host computer system is further configured to use the persona records and the relationship types to construct a family tree, receive a request from a user to provide a family tree, and send a file comprising at least a portion of the family tree to the user.

BRIEF DESCRIPTION OF THE DRAWINGS

[0013] A further understanding of the nature and advantages of the present invention may be realized by reference to the remaining portions of the specification and the drawings wherein like reference numerals are used throughout the several drawings to refer to similar components. Further, various components of the same type may be distinguished by following the reference label by a dash and a second label that distinguishes among the similar components. If only the first reference label is used in the specification, the description is applicable to any one of the similar components having the same first reference label irrespective of the second reference label.

[0014] Fig. 1 illustrates a genealogy investigation and documentation system according to embodiments of the invention.

[0015] Fig. 2A illustrates a method of genealogy investigation that may be embodied in the system of Fig. 1.

[0016] Fig. 2B illustrates one example of the process of relationship correlation in greater detail.

[0017] Fig. 2C illustrates an exemplary consolidated person page according to embodiments of the invention.

[0018] Figs. 3A-3Q illustrate a detailed example of a record consolidation process according to an embodiment of the invention.

[0019] Figs. 4A-4D illustrate a series of display screens that a user may encounter when using an embodiment of a system according to the present invention.

DETAILED DESCRIPTION OF THE INVENTION

[0020] Embodiments of the present invention provide systems and methods for genealogy investigation. In some embodiments, the present invention comprises systems and methods for receiving data from any combination of a number of sources and storing the data as records in various standardized and/or proprietary formats. Records may correspond to persons, either living or deceased, information about the persons, and relationships among them. In some embodiments, the records are used to produce family trees, either in response to a request from a user or continuously as new data is received. Thus, embodiments of the present invention provide systems and methods for taking data identifying a specific individual from any source and in any format, converting it into a common format (a persona), identifying what parts of that data may define relationships with other persons on which data is available, and processing the various data elements (persona) into pedigrees, without regard to whether any of the data elements have been so combined prior to that processing, whether in GEDCOM or any other family history format.

[0021] In contrast to previously-known "open" family tree systems, embodiments of the invention described herein treat new information merely as additional data. This is the case whether the data comes from random users or from highly reliable records systems. No

information is categorically deemed "correct" and thus does not "overwrite" data provided by others. Many previously-known systems sufferer from a bias in favor of the most recently submitted data, resulting in confusion when two data sources disagree. Those skilled in the art will appreciate this problem by realizing how different users with access to the same open system may alternatively and continuously overwrite each other's entries, especially if they disagree on some aspect of a family tree or a person.

[0022] Also in contrast to previously-known systems, embodiments of the invention described herein are "data-centric" as opposed to "tree-centric." This means that embodiments described herein collect information and store the information as data records that represent tree elements (*e.g.*, nodes and relationships). The elements, however, are not conclusively linked together and the information therein is not deemed correct, but instead the information is used to infer relationships and attributes when the likeliness exceeds a threshold. As a result, new information may either strengthen, diminish, or not affect an existing inference of a relationship or information about a person. Conversely, many previously-known systems collect data using a tree structure. New information is added only by linking off of existing trees or starting a new tree. The tree structure is the essence of the data gathering process. If a user adds new information by creating a seemingly incorrect relationship, the situation is corrected only by dissolving the relationship. Once the relationship is dissolved by a subsequent user, the previous user's interpretation of information that lead to the perceived existence of the relationship is gone.

[0023] As used herein, the term "tree" or "family tree" will refer to a hierarchical structure that links generations in parent-child relationships. It should be understood that a tree may be as simple as one parent and one child or as complex as the theoretical "single family tree" that links all individuals. Thus, any specific tree may be a part of another tree; the two may overlap, or one may completely include the other.

[0024] Trees are made up of nodes and relationships. Nodes represent persons, either living or dead. Relationships exist between nodes and represent real life relationships between the persons represented by the nodes. Relationships include mother, father, child, spouse, sibling, self or same as, and the like.

[0025] As used herein, "persona" will be understood to mean an instance of a person and a "persona record" is a data record of information from a single source that describes the person. Many different persona records may represent any given persona.

[0026] A persona may have one or more "assertions," which are presumptive truths about the persona. An assertion (or "inference") may be an event such as birth, death, draft registration, and the like. An assertion also may be an attribute such as name, occupation, race, hair color, fingerprint, DNA, and the like. An assertion may become such because an individual believes it to be true. As will be described, however, an individual or the system described herein may generate an assertion based on a review of other information. For example, based on a comparison between records, an inference of a relationship or an attribute may result. Assertions, however, may be rejected by users and/or may be overcome by new information.

[0027] "Primary source" or "primary source data" refers to a source of non-compiled genealogy information or the data therefrom. For example, a census database is a primary source, as is a news paper.

[0028] Having described embodiments of the invention generally, attention is directed to Fig. 1, which illustrates an exemplary system 100 according to embodiments of the invention. The system includes a host computing system 102 and a network 104 through which the host computing system communicates with user computers 106, tree databases 108, and records databases 110. The host computing system 102 may include a processing system 112, a storage system 114, a web server 116, administrative computers 118, and the like. The host computer system 102 includes software that programs it to perform the methods described herein.

[0029] The various elements than make up the host computing system 102 may be co-located at a single facility or distributed across a geographic area. The processing system 112 of the host computing system 102 may be any suitable computing device, or combination of devices, that are programmable to carry out the functions of embodiments of the present invention. Examples include mainframe computers, workstations, servers, personal computers, laptop computers, and the like. The storage system 114 may be any storage device or combination of storage devices. Examples include a server, a database, or the like, or any other type of storage arrangement, and may include magnetic, optical, solid state memory, and/or the like, or any other type of storage medium. The web server 116 may be any server capable of providing a web-like interface to a network, either internal or external. The administrative computers 118 may be any computing devices capable of providing administrative users access to the operations of the system.

[0030] The network 104 may be wired or wireless, and may include the Internet, a virtual private network, a local area network, a wide area network, and/or the like. The user computers 106 may be any computing devices capable of accessing the host computing system 102 via the network 104.

[0031] The tree databases 108 and records databases 110 may be any storage devices and/or computing systems mentioned above with respect to the host computer system. Tree databases 108 and records databases 110 also may be non-electronic primary sources. These databases may include public records databases, primary sources, commercial genealogy databases, private databases, and the like. For example, the tree and records databases may comprise any of the following: Ancestry World Tree, Social Security Death Index, World Family Tree, birth certificate, death certificate, marriage certificate, draft registration, veterans, property records, census, motor vehicle, and the like.

[0032] Those skilled in the art will appreciate that the foregoing is but one example of a system according to the present invention. Other systems are possible.

[0033] Attention is directed to Fig. 2A, which illustrates a first method 200 according to embodiments of the invention. The method may be implemented in the system 100 of Fig. 1 or other suitable system. It is to be understood that the method 200 is merely exemplary of a number of equivalent methods according to embodiments of the invention all of which are within the scope of the present invention. Equivalent methods may include more, fewer or different steps than those described herein, as is apparent to those skilled in the art in light of this disclosure.

[0034] The method 200 begins at block 202 wherein a host computing system, such as the system 102 described above, receives data. The data includes assertions relating to one or more personas. Assertions may include: first, middle, and last names, name prefixes (Sir, Mr., Dr. Mrs., and the like) and/or name suffixes (Sr., Jr., III, J.D., and the like); addresses; birth dates; birth places; death dates; death places; spouse names; children names; sibling names; relationships; and the like.

[0035] Data may be received in any or a number of forms. For example, data may be in the form of a family tree or in the form of records representing individual persons. In some examples, data is received as a GEDCOM file. In other examples, data is taken from indexes of primary source records such as census and vital records. Other examples are possible, including data being received in a combination of the aforementioned forms.

[0036] Data may be received from any of a number of sources. In some examples, data is received from databases such as the Ancestry World Tree Database, the World Family Tree Database, the 1930 Mini-Tree Database, and the like. In other examples, data is received from records databases such as birth records databases, death records databases, marriage records databases, census records databases, draft card databases, and the like. In other examples, data is received from individual users as either trees or individual records. In fact, potential sources include all census records (federal, state, and local) for any country, user submitted family tree data, death indexes such as SSDI for the US or Civil Registration in the UK, newspaper obituaries, various sources and forms of vital records, the Family Data Collection, military and military pension records, and/or any database that has names, dates, places and/or relationships. Other examples are possible, including data being received from any combination of the foregoing.

[0037] At block 204, data is stored as individual records. Records may include persona records, relationship records, and the like. This process involves evaluating the data and standardizing (or normalizing) its format. Many examples of this process exist, several of which will be described in more detail hereinafter. Generally, however, each record represents data from a single source and an individual person may be represented by many different records. Thus, unlike many previously-known genealogy investigation tools, embodiments of the present invention do not necessarily assume new data to be the most accurate data and use it to overwrite existing data. In most embodiments of the invention, each time data is added, it is stored as at least one new record. In a specific example, name, birth, birth place, death, and death place are stored in a record in an "individual nodes" database, and, if the data indicates a relationship, the related names and the relationship type are stored as a record in an "individual links" database. If the data includes other information, this information is stored in an "other data" database in some embodiments.

[0038] At block 206, one or more individual node records are compared. The comparison may operate on any or all of the information in the records and may use methods know to those skilled in the art or methods that are apparent in light of this disclosure. In some cases, the comparison includes factors that account for the reliability of the source. For example, public records may be considered more reliable than user-submitted data. The comparisons also may include adjustments based on other records. For example, if a draft registration exists for an individual, a birth certificate indicating the person was born only five years prior to the registration date is likely not for the same person. Many such factors may be included.

In a specific embodiment, each comparison between two individual node records results in a factor P(s) that quantifies the likelihood that the two records represent the same person. If P(s) is greater than a predetermined threshold, the two records are provisionally determined to represent the same person. This process may be referred to as "individual correlation."

[0039] Properly correlating all individual records theoretically requires comparing every individual record to every other individual record. This process, however, quickly may become an overwhelming task given the possible number of records. Thus, the process may be simplified in any of a number of ways. In a specific example, individual correlation may be simplified using, for example, a surname index to partition data into groups based on surname. The comparison process may be further simplified using, for example, a sort on first name, birth date, or other relevant data within the individual record. The partitioning process will be explained in more detail hereinafter.

[0040] Following individual correlation, at block 208 those records that have been determined provisionally to represent the same person (*i.e.*, "same as records") undergo "relationship correlation" as will be described. In a specific example, the individual links records relating to the same as records are consulted to determine whether parent relationships exist for each. If so, the respective parent records are compared to one another, if the comparison was not previously completed during individual correlation. Each comparison results in a factor, P(f) that represents a comparison of the father records, and a factor P(m) that represents a comparison of the mother records. The P(s), P(f), and P(m) factors are then collectively used in the following formula to calculate P(s|f,m) representing a revised likelihood that the two same as records relate to the same person:

$$P(s \mid f,m) = \frac{P(f)P(m)P(s)}{P(f)P(m)P(s) + P(f')P(m')P(s')}$$

Where P(s')=1-P(s); P(f')=1-P(f); and P(m')=1-P(m). If P(s|f,m) exceeds a pre-determined threshold, then the two same as records are deemed to relate to the same person. This specific example of relationship correlation is shown graphically in Fig. 2B. It is to be understood, however, that other algorithms are possible, including ones that encompass more generations or work from ancestors to descendants, rather than from descendants to ancestors.

[0041] At block 210, records are consolidated into person pages. Person pages comprise records of consolidated information about a person and may include assertions, alternative assertions, relationships, alternative relationships, sources of the information used to compile

the person page, and the like. This involves consolidating all information from same as records into a single person page, and creating a single person page for unique records. One specific example of a person page 230 is illustrated in Fig. 2C.

[0042] At block 212, a request is received from a user to display a family tree. The request minimally includes a name of a person; however, in most instances, at least one additional piece of information about the person may be required. The additional piece of information may be an assertion about the person (*e.g.*, birth, death, birthplace, death place, and the like).

[0043] At block 214, a file is constructed using the information provided in the request. The file comprises assertions about the person identified by the requester, and a family tree using the person as the root. The information is compiled by locating a person page relating to the person, then using the person page to locate other person pages related to the person. Alternative relationships also may be included. The file also may include scores relating to the likelihood that assertions and relationships are correct. The scoring process for relationships was described above; assertions may be similarly scored. At block 216, the file is sent to the user.

[0044] In some embodiments the user is given the opportunity to "drill down" to more detailed information about someone or something in the file. In response, the additional information is located and sent to the user. In some embodiments, this information is located in the original person page or a related person page. For example, the user may be able to navigate up a family tree by selecting children of the root and having a new tree generated based on the child as a root. In other embodiments, responding to the request involves selecting information from the records in the other data database. Many such examples are possible. The drill down process is shown as block 218.

[0045] In some embodiments the user is provided the option of selecting among alternatives. If provided and the user does so, the tree may be updated based on the selected alternative. In some embodiments, the user's selections are saved for the next time the user access the same tree. The iterative process of selecting and storing alternatives is shown in as block 220.

[0046] In some embodiments, the user is given the opportunity to provide information. The information may comprise one or more digital pictures, files of text (*e.g.*, a journal of a person in the requested tree, or a note about what a user knows about the person or about the sources used to evaluate information), and the like. This information may be made available

to other users. The user also may submit genealogy data. User-submitted genealogy data is received, stored, and processed as described above. The receipt of user information is shown as block 222.

[0047] The foregoing process may be repeated periodically or continuously as new data is received. In some embodiments, a records update process takes place in batch mode. In other embodiments, the process takes place each time new data is submitted. In still other embodiments, the update process is a combination of batch and continuous and may depend on the source from which the data originates.

[0048] As new data is added to the system, probability factors relating to assertions about personas and links between personas may change. Thus, a family tree originating from the same root and presented to a user on subsequent visits may be different. This may be handled in a number of ways. In one embodiment, the user is presented with the new information upon re-accessing the system. The user then may be presented with a summary of the changed inferences and given an opportunity to accept, partially accept or reject the resulting effect on the user's family tree. In other embodiments, the information shows up as an alternative selection and the user may select among the alternatives. In still other embodiments, the system generates a message, such as an email or a list of changes on a web page, that is sent to affected users when new calculations are made that affect their trees. The options then may be presented to affected users when they next access the system. Other embodiments use a combination of the foregoing. The process of notifying users regarding updates is shown as block 224.

[0049] Those skilled in the art will appreciate that the software to implement the method described above and any variation on it may be coded in most any programming language. In a specific embodiment, however, XML is used. In other embodiments, however, XML is used to represent the data, the code to correlate and consolidate is written in JAVA and C++, and the code to display the persona to the user are is written using HTML, JavaScript, and the .NET framework. Additionally, a relational database is used to manage the data at various points in the process. The code may reside on one or more computing devices that cooperate to perform the methods described above.

[0050] Attention is directed to Figs. 3A-3Q, which illustrate a more specific example of the process of receiving, storing, and analyzing genealogy data. For example, Figs. 3A and 3B depict data being received from the Ancestry World Tree database. The data exists as one or

more GEDCOM files 302. The data is read using a data extractor 304, which may be specifically designed to extract data from a specific data storage environment. Through a data scrubbing process 306, the data is parsed and evaluated. This may involve assessing its completeness, accuracy, or other characteristics. Data whose utility or accuracy falls below a pre-established threshold is rejected to an AWT threshold failed file 308. The remaining data is stored in one or more records in specific databases. These include an individual nodes database 310, an individual links database 312, an other data database 314, and a surname index 316. The individual nodes database 310 stores individuals and core data (birth and death dates) as well as the source of the data. The individual links database 312 stores links between individuals and the type of link. The other data database 314 stores information not critical to the data evaluation and relationship analysis processes. The surname index 316 stores surnames and counts of surnames. Particular uses for each of the databases will be described in more detail below. Fig. 3B more clearly illustrates the placement of specific data from GEDCOM files into records in these databases.

[0051] As shown in Fig. 3B, a unique record is created in several of the databases for an individual entry from a GEDCOM file. Names, birthdates, and deathdates for each individual go into records in the individual nodes database 310. Names, comments, and sources go into records in the other data database 314. Relationships types and the related individual names go into records in the individual links database 312. Although not shown, surnames go into the surname index 316 along with a count of the number of records in which the surname exists.

[0052] Figs. 3C and 3D illustrate a specific example of the process for extracting data from the AWT database. Fig. 3C shows three different GEDCOM files 302. At this point, no conclusions are reached regarding whether the individuals identified in the three different GEDCOM files are related. As shown in Fig. 3D, each instance of a name results in a separate record in the individual nodes database 310. Entries in the records identify the source of the data (DB) and create a unique ID for the data (ID). Other entries include name, birth, birth date, death, and death place. Of course, in other examples other data could be included in the records. Each instance of a relationship among individuals results in a record in the individual links database 312. Each record includes links that identify the source for the data (DB1, DB2), the record identifier from the individual nodes database 310, and the relationship. Each unique surname results in a record in the surname index 316, and a record in the surname index counts the number of occurrences of the surname.

[0053] Fig. 3E illustrates a data extraction process for a census database (1930 US Federal Census). The data resides in census source files 320. The data is extracted using an extractor 322 that may be specifically designed for extracting census records. The data is then stored as records in an individual nodes database 324 relating to the census and as records in an individual links database 326, also relating to the census. Note that a data scrubbing process is not shown. It may be the case that some source data is acceptable without scrubbing. The absence of a surname index indicates that some source databases do not contribute to surname counts.

[0054] Figs. 3F and 3G illustrate a specific example of a data extraction process from a census database (*e.g.*, the 1880 US Federal Census). Fig. 3F illustrates data in a specific census record, and Fig. 3G illustrates the placement of the resulting data in the individual nodes database 324 and the individual links database 326.

[0055] Fig. 3H illustrates a data extraction process for data from a social security death index (SSDI-Social Security Death Index) database. The data exists in individual files 330 and is extracted using an extraction process 332 that may be unique to this database. The data is then parsed and stored in an individual nodes database 334. In this example, because the source does not include relationships, no entry results into an individual links database. As was the case with the census database extraction process, no data scrubbing is used and no entries are made in a surname index.

[0056] It should be noted that the three data extraction examples just described are merely exemplary. Many other such examples are possible and apparent to those skilled in the art in light of this disclosure.

[0057] Continuing with the example, attention is directed to Figs. 3I and 3J, which illustrate a process of correlating individual records. In this process, individual records from each of several individual nodes databases 310, 324, 334, 342 are compared to each other using an individual correlation function 344 to determine if the records relate to the same individual. Individual records whose data is identical or nearly identical when compared (*i.e.*, individual correlation above a threshold) are stored in a same as nodes database 346 and are presumed to identify the same individual. As shown in Fig. 3J, the records in the same as nodes database 346 include the person names and record identifiers for the related records as well as a score that represents the degree to which the records are similar.

[0058] To simplify the comparison process, the individual records may be partitioned into smaller groups. In this example, the surname index 316 is used, together with a surname partition function 340 to partition data into manageable pieces. Because surnames for the same individual may be spelled slightly differently, a phonetic algorithm such as double metaphone, SOUNDEX, and/or the like may be used to keep similar names in the same partition even if they are spelled differently. The process then may be further simplified by sorting a partition on, for example, first name, birth data and/or year or other relevant data. Records within a partition and/or within ranges in the partition are compared to each other, thus significantly reducing the total number of comparisons that must be made.

[0059] The individual correlation process discussed immediately above may fail to identify records for individuals that completely changed their name. To avoid the problems this may cause, related records may undergo individual correlation after relationship correlation. Thus, two records for the same woman who changed her name at marriage may be identified once her father is identified if, for example, her first name and birth date are the same in the two records in which her last name is different.

[0060] Fig. 3K illustrates a specific example of the individual correlation process using the AWT individual nodes database 310 and the census individual nodes database 324 created earlier in the example. The comparison based on surnames results in a correlated individuals list 350. In this simplified example, the correlated individuals list 350 only includes entries based on the name "John William Jefferson." From the individual nodes database 310, a comparison of NodeID 1 to NodeID 2 results in an entry in the correlated individuals list 350 identified as Corr ID 1. The entry includes the source (DB1, DB2) and record ID (ID1, ID2) for the compared records and the score that the comparison generated. In the case of Corr ID 1, the comparison resulted in a score of 0.8. This is because the death place differs between NodeID 1 and NodeID 2 of the individual nodes database 310. A comparison between NodeID 2 and NodeID 3 from the same database, however, resulted in a score of 1.0 as can be appreciated from Corr ID 3 in the correlated individuals list 350. The remaining entries in the correlated individuals list 350 result from other entries based on the name "John William Jefferson."

[0061] Figs. 3L and 3M illustrate a further refinement of the correlation process based on relationships. The process once again uses the surname index 316 and a surname partition function 360 to evaluate data stored in the individual links databases 312, 326, 362. The data

is extracted into a relationship correlation function 364 and the records identified as being related to same as nodes are compared. The comparison updates the scores calculated previously in the individual correlation process. Thus, the scores in the same as nodes database 346 may be revised base on the comparisons. Fig. 3N illustrates a continuation of the specific example developed thus far.

[0062] Fig. 3N relates only to the record identified by Corr ID 1 in the correlated individuals list 350. The initial comparison during individual correlation of records ejerrer-I012 and a14243-I9571 resulted in a score of 0.8. Comparing the corresponding parent records for these two records, however, results in a perfect match in both cases, a score of 1.0. This may be seen by returning to Fig. 3K and comparing NodeIDs 4 and 5 and NodeIDs 7 and 8 of the individual nodes database 310. Thus, the score for Corr ID 1 of the correlated individuals list 350 may be revise upward to 1.0, representing a combination of the three comparisons. Similar relationship comparisons are used to revise the scores for the remaining records.

[0063] Figs. 3O and 3P illustrate a continuation of the process in which records identified to be the same person are consolidated. Records from the individual nodes databases 370 (which may include the AWT individual nodes database 310) and records from the same as nodes database 346 are input into an individual consolidation process 372. The output from the individual consolidation process 372 is a record in a person pages database 374 for each group of related individual records. Thus, at the conclusion of the process, a person page exists for each group of individual records from a multitude of different sources, the records determined to have been related by calculating a score based on a comparison of the individual records then adjusting the score by comparing records linked to the source records. If the score is above a pre-determined threshold, then the records are presumed related. A final consolidation for "John William Jefferson" is illustrated in Fig. 3Q.

[0064] In Fig. 3Q, the records relating to "John William Jefferson" from the correlated individuals list 350 are condensed into a record in a persons database 380. A person page 382 includes data from the source records and lists alternative information where comparisons did not result in perfect matches. The person page includes the relevant information from the original records in the individual nodes and the individual links databases as well as the data sources. Some embodiments could also include scores for each assertion and relationship. As emphasized previously, although some data may be

disregarded for various reasons because it does not exceed a threshold for accuracy or for other reasons, no data is overwritten and therefore lost in the process. A user performing a genealogical investigation is presented with a summary of the most relevant data and may further evaluate its utility. The user is not forced to accept data that someone else has deemed accurate. The user may view alternate data to determine what he or she believes to be most accurate. The user may also later change his or her mind and choose a different set of alternate information. No information is lost in any of this analysis and choosing of data.

[0065] The foregoing example depicted in Figs. 3A-3Q will be understood by those skilled in the art to be non-limiting and merely illustrative of a process for receiving and parsing data from one or more data sources. Similar processes may operate to consolidate relationships and even entire family trees, both of which are included within the scope of embodiments of the present invention and the claims that follow.

[0066] Attention is directed to Figs. 4A-4D, which illustrate a series of screen displays that depict a user interface from a user computer to the host computer system. Fig. 4A depicts a first display screen 400 showing ancestry information about "Ruth Pabodie," the person selected for analysis by the user. The display screen 400, as with the display screens to be described hereinafter, may be displayed for the user in a browser environment, for example. In another example, the display screens may be generated by client software operating on the user's computer. Many other examples are possible. The display screen 400 includes a personal information area 402 listing information about the root person such as birth and death information, spouses, and children. Conveniently, listed information may serve as a hyperlink to more detailed information. The display screen also includes a family tree 404. The family tree depicted in this display screen 400 goes back three generations from the root person, listing Ruth Pabodie's parents, grandparents, and great grandparents. Each person in the tree may be selectable as a hyperlink. An additional information section 406 provides hyperlinks to other resources relevant to the root person. This may include user-submitted information, source records, newspapers from the root person's birth and death dates, and the like.

[0067] In some embodiments, attention symbols 408 are used to indicate the presence of alternatives relating to the information marked by the attention symbol. In this example, Ruth Pabodie's father is marked by a attention symbol 408. By selecting the attention symbol 408 next to Ruth's father, the user is presented with the display screen 410 of Fig. 4B.

[0068] The display screen 410 of Fig. 4B includes an alternative father selection area 412 having three alternatives. In this example, three records were found that could be related to Ruth as her father. Rather than force the user into using the most likely alternative (the one marked with an asterisk 414), this embodiment of the present invention allows the user to view the data and make a selection using the select buttons 416. Once the user has made the selection, or if the user chooses not to make a selection, the user may select a done button 418 to return to the previous display screen 400. Fig. 4C illustrates a similar display screen 420 for selecting among alternative birth records for Ruth Pabodie. This process was described above with reference to block 220 of Fig. 2. In some embodiments, a different symbol replaces the attention symbol 408 to indicate that the user has chosen among alternatives.

[0069] Users may also view the records associated to each of the conflicting data references by clicking on a hyperlinked number or list of source document types to view the records or sources which provided the conflicting data. This will better inform the user where the information came from and allow them to make a more informed decision about which conflicting data may be correct. Users' choices of which alternative data they believe to be correct may also be logged in the system as votes. These votes may then be tallied and used to inform the system of which choice users thought was more likely correct. This voting may then be used to change which piece of alternative data the system believes to be most likely.

[0070] As described above with reference to block 224 of Fig. 2, if new information changes inferences prior to a subsequent visit by the user, attention symbols 408 may appear in new places and/or replace symbols showing that the user has selected among alternatives.

[0071] Attention symbols may also be used to denote which nodes have new messages, comments, pictures, stories, or other new or modified data. Attention symbols may also be used to help a user locate nodes which are missing key data such as birth date, death place, etc.

[0072] Returning to Fig. 4A, a details link 422 allows the user to drill down into more detail information about a subject, in this case Ruth's personal information. By doing so, the user is presented with the display screen 424 of Fig. 4D. This process was described in more detail with respect to block 218 of Fig. 2.

[0073] Returning to Fig. 4A, the absence of specific information for a root person may be indicated with brackets 426, as is the case for the day and month that Ruth Pabodie married.

[0074] The foregoing display screens are merely exemplary of display screens that may be used in connection with embodiments of the invention. Other embodiments may include more, fewer, or different display screens, as is apparent to those skilled in the are in light of this disclosure.

[0075] Having described several embodiments, it will be recognized by those of skill in the art that various modifications, alternative constructions, and equivalents may be used without departing from the spirit of the invention. Additionally, a number of well known processes and elements have not been described in order to avoid unnecessarily obscuring the present invention. For example, those skilled in the art know how to arrange computers into a network and enable communication among the computers. Accordingly, the above description should not be taken as limiting the scope of the invention, which is defined in the following claims.